DOCUMENT RESUME

ED 254 321 PS 014 938

TITLE	Path-Referenced Assessment for Head Start Children. The Head Start Measures Project: Executive Summary.
INSTITUTION	Arizona Univ., Tucson. Center for Educational Evaluation and Measurement.
SPONS AGENCY	Administration for Children, Youth, and Families (DHHS), Washington, D.C.
PUB DATE	Jun 84
CONTRACT	HHS-105-81-C-008
NOTE	56p.
PUB TYPE	Reports - Research/Technical (143) -- Reports - Descriptive (141)
EDRS PRICE	MF01/PC03 Plus Postage.
DESCRIPTORS	*Achievement Tests; Early Childhood Education; General Science; Language Acquisition; Mathematics Skills; Measurement Techniques; *Preschool Tests; Pretesting; Reading Skills; Social Development; Spanish; *Test Construction; Test Interpretation; Test Items; Test Length; Test Reliability; Test Validity; Visual Perception
IDENTIFIERS	*Path Referenced Tests; *Project Head Start; Science Skills

ABSTRACT

The Head Start Measures Project was a 3-year study to develop a set of measures designed specifically for Head Start children. The measures are based on a path-referenced approach to assessment, in which children's performance is described in terms of their position along paths of development. A path is defined as a sequence of skills within a content area that is ordered by difficulty. A path-referenced test score not only indicates what the child has achieved but also details the skills the child is likely to master as developmental progress continues. The result of the project is the Head Start Measures Battery (HSMB), consisting of six scales: Language, Math, Nature and Science, Perception, Reading, and Social Development. There are versions for Spanish-speaking and English-speaking children. In a brief and nontechnical fashion, this report summarizes aspects of the project described in more detail in other publications. Chapter I describes the background of the Head Start Measures Project and the path-referenced approach to assessment. Chapter II describes the HSMB, the 1982-83 field test results, and the uses of the measures. Chapter III describes the development and evaluation process, the psychometric properties of the measures, and the results of research relating program characteristics to achievement. Chapter IV describes a current pilot project, which involves the dissemination of the measures and their use by a sample of 30 Head Start programs. (CB)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

Prepared for the Head Start Bureau • Administration for Children, Youth and Families • Office of Human Development Services • U.S. Department of Health and Human Services.

ED254321

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

EXECUTIVE SUMMARY

The Head Start Measures Project

Path-Referenced Assessment For Head Start Children

PS 014938

By The Center for Educational Evaluation and Measurement
The University of Arizona • Tucson, Arizona
John R. Bergan, Project Director
Allen N. Smith, Government Project Officer

Contract No. 105-81-C-008

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Kathleen M. Silvers
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Prepared for the Head Start Bureau o Administration for Children, Youth and Families o Office of Human Development Services o U.S. Department of Health and Human Services

Executive Summary

THE HEAD START MEASURES PROJECT

PATH-REFERENCED ASSESSMENT

FOR HEAD START CHILDREN

June, 1984

By The Center for Educational Evaluation and Measurement
The University of Arizona o Tucson, Arizona
John R. Bergan, Project Director
Allen N. Smith, Government Project Officer
Contract No. HHS-105-81-C-098

This Executive Summary was prepared pursuant to Contract Number HHS-105-81-C-008. The statements and conclusions contained herein are those of the Center for Educational Evaluation and Measurement, University of Arizona, and do not necessarily reflect the views of the sponsoring agency.

Table of Contents

	Page
Chapter I: BACKGROUND AND APPROACH	5
Introduction	5
The Path-Referenced Approach to Assessment	6
Chapter II: OVERVIEW OF THE HSMB	8
The Head Start Measures Battery	8
The HSMB was Developed for Head Start Children	8
Content of the HSMB	9
Administration of the HSMB	12
The HSMB is for English and Spanish Speaking Children	12
Results of the 1982-83 Field Test	13
Significant Psychometric Features	14
Scoring and Reporting Services	15
Scores and Profiles for the HSMB	16
Using HSMB Results to Plan for Learning	18
Implementing the HSMB	19
Chapter III: MEASURES DEVELOPMENT AND EVALUATION	20
THE MEASURES DEVELOPMENT PROCESS	20
Path-Referenced Measures Construction Technology	20
Cultural and Linguistic Considerations in Test Construction	22
FIELD TEST AND RESULTS	23
The National Field Test Sample	24
Data Collection	27
Achievement Results	29
Psychometric Properties of the HSMB	30
Program Variables Associated With Achievement on the HSMB	40
Chapter IV: DISSEMINATION AND UTILIZATION PILOT STUDY	45
Study Design	45
Development and Production	46
Participating Sites and Children	47
Implementation	47
Evaluation	48
REFERENCES	50

List of Tables and Figures

Table

1 Ethnic Composition of National Field Test Sample

2 Age Ranges of Children in National Field Test Sample

3 KR-20 Reliability Coefficients for the Six Scales of the HSMB

4 Mean Number and Percent of Spanish Items Judged Adequate by Reviewers

Figure

1 Math Scale Profile Report: Counting Subscale

2 Pre- and Posttest Path Scores on the Head Start Measures Battery - National Field Test, 1982-83

3 Percentage of Teachers who Reported Teaching Selected Language Skills - Spring 1983

4 Percentage of Teachers Who Reported Teaching Selected Math Skills - Spring 1983

I

BACKGROUND AND APPROACH

Introduction

The Head Start Measures Project was a three-year (1981-1984) study funded by the Administration for Children, Youth and Families (ACYF), Office of Human Development Services, Department of Health and Human Services. The purpose of the project was to develop a set of measures designed specifically for Head Start children. ACYF felt that a battery of measures was needed that would assist program administrators in facilitating children's cognitive and social development. A "path-referenced" approach to assessment was created to measure children's growth in six areas. The measures were field tested at three phases in their development. In addition, data were collected on several program characteristics in order to examine their relationship to achievement on the measures. The result of the project is the Head Start Measures Battery (HSMB) consisting of six scales: Language, Math, Nature and Science, Perception, Reading, and Social Development.

This report contains four chapters. Chapter I describes the background of the Head Start Measures Project and the path-referenced approach to assessment. Chapter II describes the Head Start Measures Battery, the 1982-83 field test results, and the uses of the measures in a relatively brief and nontechnical fashion. It is designed for readers who are interested in an overview of the major features of the assessment system. More detailed information can be found in the **HSMB Examiner's Manual**. Chapter III describes the development and evaluation process during which the measures were created and refined, the psychometric properties of the measures, and the results of the research relating program characteristics to achievement. More detailed information on the development and evaluation process and the properties of the measures can be found in the **HSMB Technical Manual**. Chapter IV describes a pilot project that is currently in progress.

The project involves the dissemination of the measures and their use by a sample of 30 Head Start programs. A complete report on the project from the period of August, 1981 to December, 1983 can be found in Bergan et al. (1984) **Head Start Measures Battery Final Report.**

The Path-Referenced Approach to Assessment

Early childhood educators have generally sought three kinds of information from assessment programs: information on the relative standing of children in norm groups, information about the attainment of educational objectives, and information about children's development associated with educational experiences. Norm-referenced assessment has provided information about relative standing. Criterion-referenced assessment has afforded information about the mastery of objectives. Unfortunately, until recently there has been no adequate procedure for assessing cognitive and social development. Consequently, Head Start programs have had to rely heavily on norm-referenced and criterion-referenced measures.

The HSMB was designed to provide information about development for use in providing for children's learning needs in Head Start programs. The measures are based on a path-referenced approach to assessment (Bergan, 1981; Bergan, Stone, & Feld, in press), that describes children's performance in terms of their position along paths of development. The path-referenced approach has several key features that distinguish it from other approaches to assessment: (1) A path-referenced measure assesses a child's position in a validated developmental sequence. (2) The child's path position reveals the specific skills that the child has mastered and the skills the child will need to master. (3) Progress is measured on a quantitative scale which indicates the skills acquired during the course of development.

A path is defined as a sequence of skills within a content area that is ordered by difficulty. The sequence of skills represents the developmental route, or path, that children are likely to follow as they master increasingly more difficult skills. Path sequences are empirically validated. Validation of the HSMB was based on data from over 1,000 Head Start children.

The fact that a path-referenced test is designed to measure development implies assessment and educational practices that are quite different from those appropriate with other assessment technologies. A developmental perspective on measurement calls for assessment devices that identify not only what the child has accomplished, but also what new learning challenges lie ahead.

The frequently used criterion-referenced strategy of describing children's achievement in terms of the proportion of skills they have mastered in a content area would not assist the teacher to target classroom activities appropriate for each individual child's learning needs, since objectives based on such results may reflect skills well below a child's current developmental level. Under these conditions, knowing the child has mastered the established objectives tells little about the developmental level of the skills the child actually possesses and underestimates the developmental level of the child. A path-referenced test score indicates not only what the child has achieved, but also details the skills the child is likely to master as developmental progress occurs.

Path-referenced assessment links gains in achievement directly to changes in developmental level reflecting the acquisition of specific competencies. In contrast, norm-referenced instruments describe progress in terms of changes in relative standing in a norm group. Further, the norm-referenced approach does not indicate the kinds of skills associated with that change. Moreover, the measure of change is not independent of the child being assessed. For example, a six-month gain for a below average child does not mean the same thing as a six-month gain for an above average child (Linn, 1981).

By providing path scores that are referenced to specific skills that are ordered to reflect developmental sequences, the path-referenced approach describes children's competencies in terms of their own past and future progress along a path. Rather than comparing children to one another, or to a norm group, path scores and profiles provide a means for viewing progress of an individual child within a developmental framework.

OVERVIEW OF THE HSMB

The Head Start Measures Battery

The Head Start Measures Battery (HSMB) is a set of six path-referenced tests designed to assess the cognitive and social development of children aged 3 through 6. The Battery consists of six scales:

- Language
- Math
- Nature and Science
- Perception
- Reading
- Social Development

The HSMB was developed for the Head Start program through funding provided by the Administration for Children, Youth and Families. Development and evaluation were carried out over a period of several years. The measures will be available for use by Head Start in the fall of 1984.

The HSMB was Developed for Head Start Children

The HSMB was designed specifically for use by the Head Start Program. In order to articulate the measures to the goals and objectives of the Head Start Program, and make them appropriate for assessing development, several sources were called upon to determine the content of the tests. First, groups of parents and Head Start staff from every region of the country provided input regarding important areas to assess. Extensive lists of objectives were generated and refined. Second, the Head Start Performance Standards was an important source for determining the content of the measures. Third, early childhood curricula were used in developing the measures. Fourth, the research literature in child development was examined. The research knowledge base made it possible to construct measures in which items were selected to reflect developmental sequences based on developmental theory. Fifth, linguistic and cultural advisors were involved throughout the test

development effort. They reviewed items under development for potential bias and reviewed the measures at each stage of development. Sixth, national advisory panels comprised of Head Start personnel and experts in child development and assessment reviewed the measures development process and the field test design.

The Head Start Measures Battery was administered during field tests to national samples of Head Start children from all regions of the country, representing several ethnic groups including Blacks, Hispanics, Native Americans, and Anglos. The purpose of the field tests was to obtain information to be used to establish the validity and reliability of the measures, and to improve the items.

Content of the HSMB

Language Scale. The Language Scale assesses skills necessary for effective communication. One major category is Meaning in Language. The understanding of events in a story is assessed, as is the ability to generate verbal explanations.

A second assessment category contains items that assess syntactic skills. Words must be sequenced to form intelligible sentences. Word endings must be varied to reflect the plural, the comparative, the possessive, and the subtleties of meaning conveyed by verb tenses. Items tap each of these abilities in English. In addition, several items assess grammatical forms unique to Spanish.

The third category addresses skills for communicating with others. This involves, among other things, the ability to vary one's language based upon what one can assume the listener knows. Another important skill assessed is knowledge of the social rules which govern interactions such as greetings, turn-taking, maintaining and changing a topic, and asking questions.

Math Scale. The Math Scale focuses on children's skills in computation and in number and measurement concepts. The first category, Working With Numbers, includes recognizing numerals, counting, adding, and subtracting. In addition, items tap children's understanding of the fact that set size remains the same regardless of the arrangement of objects within it. The use of numbers to indicate position,

for example, first, second, third is also assessed.

The second category, Working With Measurement Units, assesses knowledge of the names and functions of measurement units. Children are asked to compare the size or sequence of length and time units.

Nature and Science Scale. There are two major categories in the Nature and Science Scale. The first focuses on content or subject matter, and the second on the processes engaged in by scientists in gathering and organizing information.

The following kinds of subject matter are assessed: earth and universe, weather and seasons, time/space, mechanics, motion, energy, substances, ecosystems, human health and anatomy, life cycles, physiology, agriculture, goods and services, transportation, and tools and construction.

The process category assesses skills in observation of the physical and action characteristics of objects and animals, and their classification based on these characteristics. Another process assessed is making inferences and explaining causal events. Knowledge of relationships, recognition and construction of meaningful sequences, and prediction of the outcome of an action are assessed in a variety of items.

Perception Scale. The Perception Scale assesses children's developing skills in recognizing shapes and in understanding the spatial relationships among objects in the physical world.

The Shape category requires constructing shapes from parts and matching shapes which have been rotated.

The Relations category has three separate subcategories. The first deals with objects in relation to each other. Items require constructing a three-part display which matches the model. A second subcategory assesses children's ability to understand that alternative points of view of a single object are possible. The third subcategory involves matching or constructing a repeating pattern. Items require building a matching pattern or completing a repeating pattern.

Reading Scale. A variety of prereading skills are assessed in the Reading Scale. They are divided into two major categories: Visual Processing and Contextual Processing.

Visual Processing includes such skills as letter naming, choosing from a group of letters one which has already been seen, identifying a corresponding upper or lower case letter, rhyming, dividing words into their component sounds, and recognizing that words (such as chalk-board) may be composed of other shorter words.

Contextual Processing on the other hand requires that the child utilize previous knowledge or experience rather than rely primarily upon the printed text. Contextual Processing tasks involve asking the child to recognize his/her own name, and to fill in an appropriate word in an incomplete sentence.

Social Development Scale. The Social Development Scale assesses the child's knowledge of social relationships. It is divided into three categories: Social Roles, Social Rules, and Feelings.

The first, the Social Roles category, taps the child's understanding of the requirements and expectations involved in the roles of the leader, the buyer and the seller, and the owner of both tangible and intangible property.

The Social Rules category assesses the child's understanding of the concept of fairness, as it relates to the allocation of rewards. The understanding of turn-taking, and of helping others and sharing is also assessed.

The third category, Feelings, asks the child to identify feelings from facial expressions and to predict the feelings of the recipient of certain actions.

Pilot data were collected during the spring of 1983 for two additional social development categories. The first assesses self concept with respect to cognitive skills and the second taps classroom social skills. Items in these categories are being incorporated into the item bank for the HSMB so that they may be used as needed in future versions of the battery.

Administration of the HSMB

The Head Start Measures Battery was designed to be easy to administer so that it can be administered by teachers. It also has been administered successfully by many paraprofessionals. Although prior testing experience is desirable, it is not necessary, since most items employ a simple 1 or 0 scoring system. A training package consisting of an **Examiner's Manual** and 1/2" video cassettes is available for self-training. In addition, a **Data Collection Training Manual** details procedures for training personnel to administer the HSMB.

Each scale of the Head Start Measures Battery is individually administered. Administration time ranges from eight to twelve minutes depending on the scale and the particular child being tested. Rules for starting and stopping the administration of specified subscales serve to shorten time as well as to avoid administration of very difficult items to children at low achievement levels and very easy items to children at high achievement levels.

Sometimes young children have knowledge that they do not demonstrate in a testing situation for a number of reasons, including lack of rapport with the examiner, or lack of understanding of what is required in the task or of the language of testing. The HSMB outlines specific procedures designed to maximize children's performance on the Battery and make the testing experience an enjoyable one. Many of the test items are preceded by demonstration and practice items that serve to show children exactly what is expected of them. Techniques for building and maintaining rapport and providing encouragement are spelled out. In field tests, children generally enjoyed the activities in the Battery and were usually eager to "play games again" with the examiner.

The HSMB is for English and Spanish Speaking Children

An important feature of the HSMB is that it allows for linguistic diversity and thus yields more accurate information about children's knowledge. When children are tested in a language they do not fully comprehend, they may miss items, not because they do not have the knowledge but because they do not understand the examiner.

Testing a child in the appropriate language or languages helps to ensure that results will provide a valid indication of what the child knows.

Approximately twenty percent of the Head Start population is Hispanic. While the majority of the Hispanic children are of Mexican-American background, Puerto Ricans, Cubans, and other Hispanic groups are also represented. One goal of the measures development effort was to produce measures capable of adequately assessing competencies for each of these groups. To this end, the measures were developed in both Spanish and English, and the Spanish versions were constructed so that they would be appropriate for speakers of different varieties of Spanish.

The Spanish and English versions of the measures are both contained in the test manuals so that it is not necessary to obtain a separate manual for each language. In most cases Spanish and English items are equivalent; however, there are certain areas of competence in one language that do not exist in the other language. In such cases, separate sets of items were developed. In the Language Scale, for example, the English version contains a subtest of grammatical structures that exist only in English and the Spanish version contains a subset of grammatical structures that exist only in Spanish.

The **Examiner's Manual** outlines specific procedures for determining whether a child should be tested in English, in Spanish, bilingually, or not at all. The procedures are consistent with those recommended for use with the Head Start bilingual curriculum models that have been developed. The HSMB is not intended for use with children whose dominant language is neither English nor Spanish because valid results cannot be expected with children who are not dominant in one of these two languages.

Results of the 1982-83 Field Test

Results of the field test demonstrate the sensitivity of the HSMB to growth made by children who participated in the Head Start Measures Project. Gains made between pre- and posttesting are presented in Figure 2 in Chapter III. The testing interval was approximately seven months for most of the measures, but only about four months for language. Data analysis indicated that the amount of

growth made by the children was educationally meaningful and cannot be attributed solely to maturation. While age was a factor related to gains, Head Start program variables also played a role in determining the levels of achievement reached by the spring.

Significant Psychometric Features

The psychometric properties of the spring 1984 version of the HSMB are discussed in detail in Chapter III. The present section briefly outlines particularly significant psychometric features of the instruments. The psychometric properties of the HSMB were established through the use of classical item analysis procedures and newer latent variable techniques described later in the summary.

- Reliability: In any assessment program, it is essential to know the extent to which a test score provides a consistent measure of a child's performance. Consistency is determined by calculating test reliability. KR 20 reliabilities ranging from .83 to .92 demonstrate the reliability of the scales of the HSMB.

- Item Discrimination: The extent to which items can discriminate among children of varying achievement levels is an important consideration in test construction. Latent trait procedures were used to calculate item discrimination values. Those items that are sensitive to different levels of ability were selected for the HSMB.

- Item Difficulty: Each scale of the HSMB contains items with a broad range of difficulty levels appropriate for children of different ages and developmental levels. Items were included only if they could be passed by children at one of more of the developmental levels for which the measures were designed.

- Item Information: Items vary in the amount of information that they provide about a child's ability. Test length can be kept to a minimum when items with high information values are used in test construction. The amount of information that items contributed to the total test score

was one criterion for selecting them to be included in the final version of the HSMB.

- Item Bias: Three methods were used to predict and detect potential item bias: judgmental review, comparisons of difficulty values, and comparisons of item characteristic curves. Four groups were studied: Blacks, Spanish-dominant Hispanics, English-dominant Hispanics, and Anglos. Analyses showed few items to be potentially biased, and no one ethnic group was favored.

- Content Validity: Cultural advisors assessed the appropriateness of HSMB content for children from different ethnic backgrounds resulting in the elimination of items with potentially biased content. Reviews by experts established the appropriateness of the content of each scale and the technical quality of the items. Additional evidence of content validity was provided by Head Start teachers. Results of the field test demonstrated that all of the items reflect skills taught in Head Start.

- Construct Validity: Several analyses were conducted that established the fact that the HSMB assesses development. Studies were conducted that empirically validated developmental sequences for each of the scales in the HSMB. These developmental sequences were based on research and theory in child development.

Scoring and Reporting Services

- Score sheets are currently available for optical scanning at The University of Arizona's Center for Educational Evaluation and Measurement. Score sheets can be developed for hand scoring by individual grantees if needed.

- Reports are provided summarizing performance and developmental progress for individual children and classes. Reports make it easy to relate a child's developmental level to the level of classroom instruction. The teacher using the reports can plan learning activities based on the child's developmental level and can evaluate progress

associated with Head Start experiences.

Scores and Profiles for the HSMB

A path score and path profile are provided for a child on each of the six scales of HSMB. The path score represents the child's position on a developmental path. It describes the child's overall level of performance on each scale and, in addition, generates a profile of the child's level of achievement on each subscale. The profile indicates what types of learning activities are easy for the child, what activities should be challenging to the child at his or her current level of development, and what activities are probably still too difficult for the child.

Figure 1 illustrates the Teaching and Skill Developmental Profile Report for the Counting Subscale of the Math Scale. Notice that the skills are ordered by difficulty. The uppermost skill is the easiest and the lowest skill is the most difficult. Note also that the path scores for each skill are partitioned into three categories: nonmastery, partial mastery, and mastery. The higher the path score, the more likely the child is to be able to master increasingly difficult skills. For example, a child with a path score of 52 is considered to be a master of the first three skills in the Counting Subscale, a partial master of the next two skills, and a nonmaster of the last two skills. On the other hand, a child with a path score of 67 is able to accurately perform <u>all</u> the skills in this subscale. Partial mastery can be interpreted to mean that the child performs the skills in some situations but not in other situations. The teacher may want to provide additional learning opportunities with these skills in order to improve the child's degree of mastery.

Teaching and Skill Profile Report - Math Scale

Site:
Center:
Class:
Date:

Mastery:
Partial Mastery:
Nonmastery:

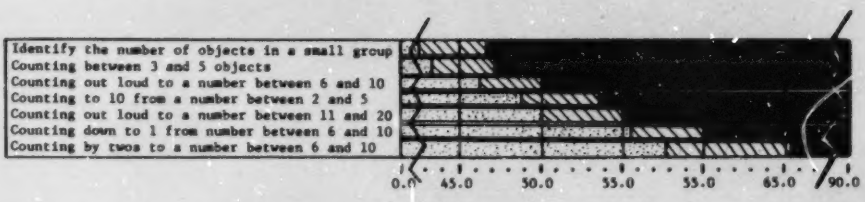

Identify the number of objects in a small group
Counting between 3 and 5 objects
Counting out loud to a number between 6 and 10
Counting to 10 from a number between 2 and 5
Counting out loud to a number between 11 and 20
Counting down to 1 from number between 6 and 10
Counting by twos to a number between 6 and 10
0.0
45.0
50.0
55.0
55.0
65.0
90.0

Figure 1. Math Scale Profile Report: Counting Subscale

BEST COPY AVAILABLE

Using HSMB Results to Plan for Learning

The measures can be very useful to educational coordinators and teachers in planning individualized and group learning activities to enhance children's development. They are also useful in assessing developmental progress and in evaluating the appropriateness of activities. Fall administration of the measures makes it possible for teachers to base long-range plans on the entering skill levels displayed by children. They are provided the necessary information to avoid the pitfalls of targeting learning experiences on skills that children already possess or of focusing learning activities at developmental levels too far above the current levels of functioning of the children. Spring administration provides an indication of the extent to which educational objectives have been achieved and, when combined with information from the fall, details developmental gains occurring during the course of the year. This information is important for subsequent educational planning. HSMB results can be used to accomplish the following:

- Describe a child's current performance level.
- Create class and program profiles,
- Revise educational goals,
- Plan learning activities at appropriate levels,
- Determine for individuals and for classes how much growth has been made over a year,
- Assess whether educational goals have been achieved.

A teacher **Planning Guide** is provided with the HSMB. The **Guide** enables teachers to relate the level of teaching planned for the child to the developmental level of the child. By using the **Planning Guide** teachers can determine whether the planned and actual teaching levels are above, approximately at, or below the child's developmental level.

Implementing the HSMB

A series of manuals accompanies the HSMB. The **Technical Manual** describes the psychometric properties of the Battery. The **Examiner's Manual** outlines procedures for administration and scoring and explains how to interpret results. The **Data Collection Training Manual** provides procedures for training and monitoring.

III

MEASURES DEVELOPMENT AND EVALUATION

The Measures Development Process

As described briefly in Chapter II, several features characterized the process of developing the HSMB. A fundamental characteristic of the measures is that they are based on child development theory and research. Such theory and research provided information on the nature of children's cognitive and social skills and the kinds of changes expected in these skills during the course of development. A second key feature of the measures is that they were designed specifically for the culturally diverse Head Start population. Systematic consideration of cultural differences and the detection and elimination of potentially biased items at several phases of measures development were the procedures used to help ensure fair assessment of children from varying backgrounds. Spanish versions of the measures were developed so that Spanish-preferring (or Spanish-dominant) children could be assessed. Another feature of the measures is that the content of the HSMB was geared toward the educational goals of Head Start programs. Therefore, extensive and systematic procedures were implemented early in the project to obtain input from the Head Start community. This section describes the path-referenced measures construction approach and the methods employed to make the HSMB culturally appropriate.

Path-Referenced Measures Construction Technology

The development of path-referenced assessment instruments for Head Start began with the construction of theoretical models of the structure of competence in each of the six areas targeted for assessment: Language, Math, Nature and Science, Perception, Reading, and Social Development. The theoretical models are referred to as "domain structure models." The domain structure model provides an organized theoretical framework for constructing a measure containing items that adequately represent a content area and, in

addition, tap a range of levels of development. Each scale of the HSMB can be thought of as providing a linear sequence that quantifies a child's position on a "path" of development. Two types of analyses were carried out to create the domain structure models: (1) content category analysis and (2) developmental structure analysis.

Content Category Analysis. Content category analysis was used to identify the types of content included in the scales. Each of the six broad areas of competence was divided into categories. For example, the area of mathematical knowledge was divided into working with numbers and working with measurement units. These categories were divided into subcategories. Working with numbers, for example, was divided into two subcategories, computation of numbers and identification of math symbols. This process of subdivision continued until a set of task strands, each including a number of tasks sharing a common goal and involving an organized set of processes directed toward attainment of that goal, was determined. For example, addition tasks share a common goal, the summing of numbers. Thus, addition tasks make up a task strand. Task strands form the building blocks from which the developmentally sequenced measures were constructed.

Developmental Structure Analysis. Following the content analysis, an analysis of developmental structure was undertaken to determine the sequence of skills within each of the six content areas. The ordering of skills is based on two factors that are specified and thus built into the items that make up a particular measure: (1) variation in the objects used to present tasks and (2) task demands.

Variations in the objects used in presenting a task may affect the way in which a task is approached by the child. In the case of addition, for some items the numbers to be added are presented in the form of concrete objects such as blocks. Under these conditions, the child could add by counting all the blocks. In other items, however, numbers are presented in the form of verbal symbols. Under these circumstances, a child might add by relying on verbal knowledge of addition facts. Thus, whether or not a task involves blocks or verbal symbols contributes to determining where in a sequence of skills the task will fall.

Another way in which the sequence of tasks was determined was through the specification of task demands. A task demand is characteristic of a task that affects the difficulty of performing the task. Task demands affect the cognitive processes necessary for task performance. For example, the addition of two numbers with a sum less than ten requires fewer steps, and hence fewer processes, than does an addition task that requires carrying. Thus, the demand of carrying influences the difficulty of the task because tasks that have this demand require more steps, and hence more processes, than do tasks that do not require carrying. In addition to affecting complexity, task demands affect the types of rules that are required for task performance. For example, when children first learn to count, they may operate under a rule that assumes that counting always begins with the number 1. As development progresses, they may replace this with a rule indicating that counting may start from any number. The latter rule is more complex because it replaces a single starting value with a whole class of values. These methods of analysis were used to order tasks within task strands, thus creating the developmental structure that is an essential feature of the path-referenced measures.

Cultural and Linguistic Considerations in Test Construction

A number of procedures were implemented to construct a battery of measures appropriate for a culturally and linguistically diverse population of children. In addition to post hoc analyses of bias, the prevention and elimination of potential bias was an integral part of the measures development process.

Several procedures were utilized during the development of items. First, procedures required the specification of cultural considerations for each item. For example, items were designed so that alternative methods of demonstrating skills among different cultural groups would be accepted, and cultural differences in item administration procedures were also considered. Second, verbal responses by the child were kept to a minimum on most of the scales and many items require non-verbal responses. Third, both Spanish and English versions of the HSMB scales were developed so that the measures could be administered to Spanish-speaking children.

The development of the Spanish version of the HSMB involved more than a simple translation of items from English into Spanish. First, there are areas of content that exist in one language but not in the other. In such cases, items were developed separately for the Spanish and English versions. Second, since Spanish-speaking children in the United States speak several regional varieties of Spanish, it was important to ensure that the Spanish words and phrases used in the measures be comprehensible to children from different regions. For Spanish terms that differ by region, alternative forms are provided in the measures. Systematic sociolinguistic survey procedures were implemented during the field test to gather information on the adequacy of the Spanish versions and the results were used to revise the measures after both fall and spring administrations.

In order to avoid potential cultural bias, cultural bias reviews were carried out before item tryouts, after the fall 1982 field test, and again after the spring 1983 field test. The reviewers assessed potential bias against Blacks, Hispanics, and Native Americans. Item characteristics assessed were content familiarity, stereotyping, relevance, cultural meaning, offensiveness, and value assumptions. Overall test characteristics assessed were visual and name representation of minority groups, ethnicity of main and secondary characters, representation of activities and cultural characteristics of various ethnic groups. Results of the reviews were used to revise the measures during the developmental process. In addition to the judgmental reviews, statistical analyses of potential item bias were carried out after fall and spring field tests and results used to eliminate items on which groups performed differently given the same overall ability level.

Field Test and Results

The national field test was conducted in the fall of 1982 and the spring of 1983 in order to try out the measures on a large sample of Head Start children and to gather data that could be used to establish the validity, reliability and other characteristics of the measures. Information from the field tests was used to revise the measures after both the fall and the spring administrations. This section

describes the sample of children who participated in the field test, data collection procedures, the amount of growth made by Head Start children between fall and spring testing, the psychometric properties of the HSMB, and results of the analyses of the relationship between selected program variables and achievement.

The National Field Test Sample

Approximately 1400 children from 19 sites were tested on the fall field-test version of the HSMB in the fall of 1982. In the spring of 1983, approximately 1100 children in 14 sites were tested on the spring field-test version. Children in the sample represented four major regions of the nation: the Farwest, Midwest, Southeast, and Northeast. Three community types were represented: rural; urban and suburban; and urban, suburban, and rural.

Table 1
Ethnic Composition of National Field Test Sample
Spring 1983

Ethnic Group	Percent in Sample	Percent in Head Start Population
Black	30	42
Hispanic, Spanish dominant	14)	20
Hispanic, English dominant	19)	
Native American	15	4
Anglo	21	33
Other	1	1
Total	100	100

Table 1 presents the percent of children in the spring 1983 sample by ethnic group. Children were selected from five major ethnic groups: Blacks, Spanish-dominant Hispanics, English-dominant Hispanics, Native Americans, and Anglos. The total percentage of Hispanic children in the sample (33%) was larger than their percentage in the Head Start population (20%). This type of sampling plan was necessary

in order to gather a sufficient amount of information on the Spanish versions of the scales. The percentage of Native Americans in the sample (15%) also exceeded their representation in the population (4%). The larger absolute numbers were needed in order to obtain information on how Native American children respond to the measures. The oversampling of Hispanic and Native American children with respect to their representation in the population necessitated the undersampling of Blacks and Anglos; however, the absolute numbers of Blacks and Anglos were adequate to gain the necessary information on their responses to the measures.

Table 2
Age Ranges of Children in National Field Test Sample
Spring 1983

Age Range	Number of Children	Percent of Children
Level I		
3-6 to 3-11	96	16
4-0 to 4-5	155	26
4-6 to 4-11	348	58
Total	599	100
Level II		
5-0 to 5-5	365	89
5-6 to 6-0	47	11
Total	412	100
Total Tested	1011	

Note. Children were 6 months younger when tested in the fall of 1982.

Characteristics of Children. Table 2 presents the number and percent of children in various age ranges for

Levels I and II of the spring 1983 field-test version of the HSMB. The ages of the children in the spring 1983 sample ranged from 3-6 to 6-0 years. While age data are only reported for the spring, it should be noted that children were about six months younger for fall testing. The younger children were all attending Head Start. During the 1982-1983 field test, each scale was divided into two levels. Level I was used with children ranging in age from 3 years to 4 years and 11 months. Level II was used with children over 5 years of age. Ninety-six percent of the children who took Level I were Head Start children, and 84% of those who took Level II were in Head Start. Of the 599 Head Start children assessed on Level I, 58% were between 4-6 and 4-11. At Level II, 89% of the children assessed were between 5-0 and 5-5. Small numbers of elementary-school-aged children were included in the sample to make it possible for the HSMB to be used to document the progress of Head Start children into elementary school.

Family Characteristics. Descriptive data were gathered on families, classrooms, and teachers of the children in Head Start classrooms. Data on family size, education, income, and occupation were gathered both from families and from Head Start program records.

- The mean number of family members for children in the sample was 5;
- Approximately half the children were from two-parent homes, and half were from single-parent homes;
- The mean annual income was $6,132;
- While there was some variation in the educational level of mothers, 362 of the 855 sampled mothers (42%) reported that the 12th grade was the highest grade they had completed;
- The occupation reported for the largest single group of mothers was service worker (59 of 151) while laborer was the occupation most often reported for fathers (85 of 257).

Staff Characteristics. A short Classroom Staff Questionnaire was administered to the teachers of all Head Start children in the sample to gather descriptive information

on education, training, and experience.

- o The mean number of years of education for sampled Head Start teachers was 13.8;
- o The mean number of months of CDA training reported by Head Start teachers was 10.34;
- o 36% of the teachers reported having the CDA credential, 29% were working toward the CDA credential; and 66% reporting having had CDA training;
- o 26% of the teachers had state teaching certificates;
- o In general the teachers in the sample were experienced with a mean of 6.8 years of experience as a Head Start teacher.

Class Characteristics. Information about the composition of classrooms was gathered through direct observation. The mean Head Start class size for children in the sample was 17.16 children. There was, on the average, one teacher, one aide, and one volunteer per classroom. In addition, there was a mean number of 1.67 children with special needs in the class and the mean number of limited English speaking children was 1.95. The mean number of limited English speaking children per class ranged from 0 in several sites to 12.64 in one site. Children with special needs and children who were limited in English proficiency and spoke a language other than Spanish were not assessed in the field test.

Data Collection

Data Collection Management. Fall data collection was carried out by Mediax Associates, a Connecticut-based research firm. Spring data collection was implemented by The University of Arizona. In both the fall and the spring, data collection was supervised by a data collection manager who supervised site managers working in the field. On-site data collection was directed by site managers, each of whom directed and monitored the activities of from two to twelve data collectors. The group of site managers included many university teaching staff and experts in child development.

Site managers attended an intensive one-week training session and were provided with a comprehensive manual of instructions. They then interviewed, hired, and trained data collectors, each of whom had to be officially designated as adequately proficient in measures administration before testing children.

Data Collector Qualifications. Data collectors who carried out the testing were paraprofessionals who met the following criteria:

- o Familiarity with Head Start and its goals;
- o Experience working with young children;
- o A high school diploma;
- o Proficiency in reading, speaking, and understanding Spanish (at some sites).

Preparation for Measures Administration. The local sample of children was selected with the cooperation of the Head Start Director. Parent permission was obtained for children chosen to participate in the project. Special needs children identified by Head Start personnel as having physical or mental handicapping conditions such that a valid assessment could not be obtained were not included in the field-test sample. In addition, children who were limited in English proficiency and did not speak Spanish were not assessed. The language(s) in which measures were to be administered to each child was determined by gathering information from parents, teachers, and aides about children's language use patterns. A language preference rating was assigned to each child and the measures were administered either in English, Spanish, or bilingually, as designated. Under field test conditions the bilingual examiners were permitted to switch languages if, in their opinion, the child's performance indicated that the language preference rating was in error and more valid results were obtained using the other or both languages.

Administration of the Measures. Measures were administered individually to each child in a location close to but separate from the classroom. Examiners followed strict guidelines about building rapport with the child, making the experience pleasant, providing positive comments,

and stopping an administration if the child became tired or disinterested.

Monitoring and Quality Control. The data collection manager and supervisors monitored site managers in the field. Site managers in turn monitored their data collectors on a daily basis. Monitoring and quality control on site involved: (1) using a Monitoring Form and Examiner Competency Rating Form to monitor each measure; (2) shadow scoring of each data collector on each measure to ensure 95% accuracy in scoring; (3) reviewing all protocols for completeness and correctness; and (4) maintaining communication to identify and solve problems. Quality control procedures implemented at the University of Arizona for incoming data included visual inspection of all protocols for: completeness; appropriate test level for child's age; proper codes for site, center, class, data collector; administration time; and language of administration. The use of a quality control computer program served to flag data collectors with problems and to assess the quality of the data files. File accuracy was maximized through visual inspection and correction of errors.

Achievement Results

The administration of the HSMB at two points in time during the year served several purposes. First, it afforded the opportunity to determine whether the HSMB is sensitive to growth made by children as a result of participation in Head Start programs. Second, it was necessary in order to establish that the difficulty level of items was appropriate for both the youngest children in the fall and the oldest children in the spring. Third, data from fall and spring administrations provide baseline data that could later be used to determine whether larger gains result when teachers use HSMB results to improve educational activities. Results of the field test demonstrate the sensitivity of the HSMB to growth made by children who participated in the Head Start Measures Project. Gains made between pre- and posttesting are presented in Figure 2. The testing interval was approximately seven months for most of the measures, but only about four months for language. Data analysis indicated that the amount of growth made by the children was educationally meaningful and cannot be attributed solely to maturation. While age was a factor related to gains, Head Start program variables also played a role

in determining the levels of achievement reached by the spring.

Psychometric Properties of the HSMB

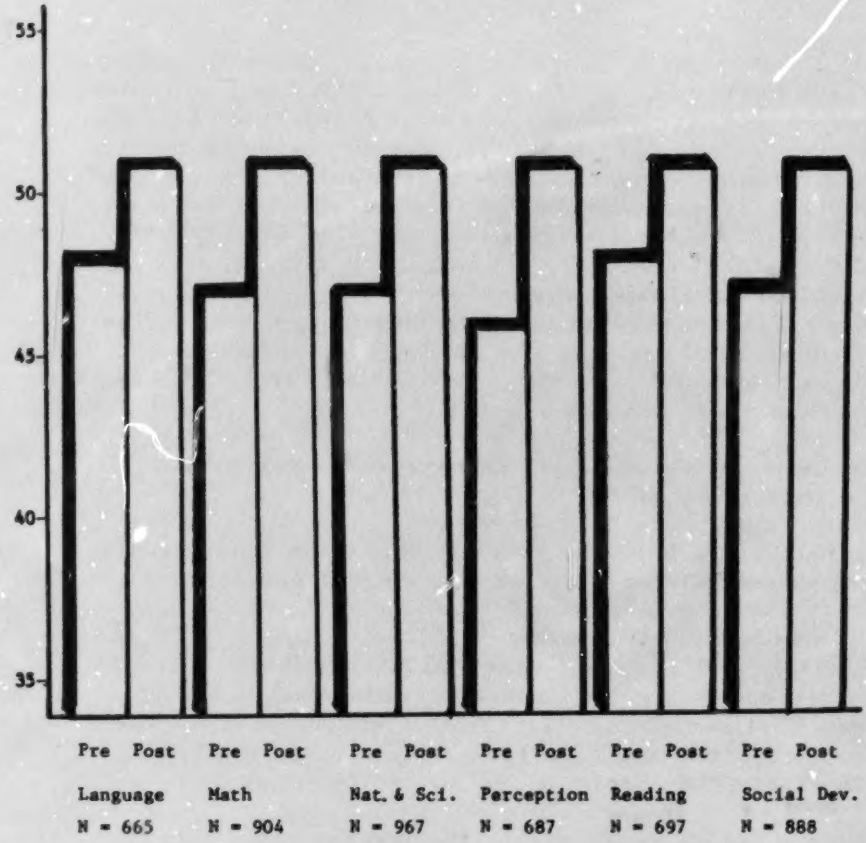
The psychometric properties of the HSMB were established using both classical psychometric procedures and newer latent trait techniques (Lord, 1980). Latent trait procedures describe a child's competencies in a given content area in terms of a continuous quantitative scale reflecting the child's level of competence with respect to the ability or trait being measured. Latent trait techniques are ideally suited to the task of measuring development. A child's score on a latent trait scale specifies the child's level of development. The score indicates the skills the child has mastered and those that must be learned in order for the child to progress.

Latent trait techniques proved to be useful in a number of ways in constructing the HSMB. In developing the battery, it was necessary to construct scales for children from three through six years of age. This required that items appropriate for older children be placed on the same scale as items appropriate for younger children. Latent trait procedures were used with each scale in the battery to place a pool of items for young children onthe same scale as a pool of items for older children.

The construction of the HSMB called for the generation of scales that can be administered in a relatively short time span. The test length required to accurately assess achievement depends on the amount of information that each of the items in the test is capable of conveying about a child's achievement level. Latent trait technology includes procedures for determining the amount of information contained in test items. These procedures were used in the construction of the HSMB to develop scales providing a maximum amount of information about children's skills with a minimum number of test items.

In order to create an adaptive measurement system capable of accommodating local program needs and a changing Head Start program focus, it was necessary to construct the HSMB scales in a manner that would allow old items to be

Figure 2. Pre- and Posttest Path Scores on the Head Start Measures Battery - National Field Test, 1982-82



BEST COPY AVAILABLE

deleted and new items to be added to the existing scales. Latent trait technology affords the capability to add and/or delete items to an existing set of scales. This feature of latent trait techniques was used extensively in the refinement of the Battery and can be used in the future to ensure that the Battery will remain up to date.

Data documenting the psychometric properties of the HSMB are presented in the **HSMB Technical Manual**. A summary of psychometric analyses is presented in the following paragraphs.

Item Difficulty. In order to accurately assess the varying skill levels represented by the Head Start population, a wide range of difficulty values is essential. For example, if the measures contained only very easy items, they would not be useful in assessing children with advanced skills. Likewise, if the measures contained only very difficult items, they would not be useful in assessing the competencies of children in the early stages of development. All of the scales in the HSMB contain an appropriate and wide range of difficulty values, indicating that they provide accurate assessment of the varying skill levels represented in the Head Start population.

Item difficulty values for the Head Start Measures Battery were calculated using classical techniques and latent trait estimation procedures. In classical test theory, item difficulty is defined as the proportion of individuals who pass an item. When proportion correct is used as an indicator of difficulty, difficulty varies with examinee ability. When an item is given to high ability examinees, a higher proportion will pass it than would be the case if the item were given to low ability examinees. This would describe the item as being easier than would be the case if the item were calibrated on a lower ability group. Latent trait estimates of difficulty, in contrast to classical estimates, do not vary as a function of examinee ability. This was an advantage in constructing the measures because it resulted in the construction of an entire scale that is independent of examinee ability.

Item Discrimination. Item discrimination refers to the extent to which an item is sensitive to the different ability levels of children. In latent trait technology item discrimination is defined in terms of the associa-

tion of changes in ability with changes in the probability of a correct response. For example, a highly discriminating item is one for which a small change in ability will produce a measurable change in the probability of a correct response. By contrast, an item that does not discriminate well between ability levels is one for which a large change in ability is required before a measurable change in response probability occurs. Items were selected for the HSMB that discriminated well for Head Start children at all ability levels. The majority of the items in the HSMB have high discrimination values indicating that they are sensitive to changes in ability.

Reliability. The reliability of the measures was assessed using the Kuder-Richardson 20 coefficient. The K-R 20 provides a measure of internal consistency based on the average intercorrelation among items. Internal consistency is important in that one would have little faith in a measure that yielded widely discrepant scores for different subsets of items all assumed to be on the same scale. The HSMB was found to have K-R 20 values ranging from .83 to .92, indicating good levels of reliability. Table 3 displays the reliability coefficient for each of the scales in the HSMB.

Table 3
K-R 20 Reliability Coefficients
for the Six Scales of the HSMB

Scale	Orig. No. of items	No. of items on Spring 1984 version	K-R 20
Language	79	62	.90
Math	114	59	.92
Nature & Science	84	51	.83
Perception	34	25	.83
Reading	57	40	.91
Social Development	68	40	.84

Item Bias. Item bias studies were conducted after the fall 1983 and spring 1984 field tests. Analyses were conducted for four groups: (1) Anglo, (2) Black, (3) English-dominant Hispanics, and (4) Spanish-dominant Hispanics. The purpose of the analyses was to determine if, after controlling for ability level, groups performed differently on an item. The three techniques used to determine if bias existed were

- judges' ratings,
- comparison of item characteristic curves (ICC's) showing the relationship between the probability of a correct response and child ability (Lord, 1980), and
- comparison of item difficulties.

Overall, the results of the investigation revealed only a small amount of bias. Moreover, there were exceptionally few instances in which the items on a scale favored any group over other groups. In those instances, the determination of bias indicated circumstances in which one group responded differently to an item than another, but was not necessarily put at a disadvantage by the differ-

ential responding pattern. For example, some items were found to be more difficult for a group when comparing students at high ability levels but easier for the group when comparing students at low ability levels. Differences of this kind tended to occur most often with older children. At each stage of measures revision, many potentially biased items were eliminated from the Battery.

Content Validity. In order to determine the potential usefulness of the measures battery in Head Start, it was necessary to address questions concerning the extent to which the content of the measures reflected the competencies that the measures were supposed to assess. The establishment of content validity was addressed through reviews of both the conceptual papers and the measures by experts external to the project. Information from these reviews was utilized during the measures revision process. Additional evidence of content validity was provided by Head Start teachers. Data obtained through use of the **Planning Guide** indicated that all of the skills assessed on the HSMB were taught by some proportion of the Head Start teachers in the sample. In the area of language, for example, about a third of the skills assessed on the Language Scale were to be taught by between 61% and 100% teachers. Another third were taught by between 41% and 60% of the teachers, and the remaining third skills were taught by between 21% and 40% of the teachers. In the area of math, one quarter of the skills assessed on the Math Scale were taught by between 61% and 100% of the teachers. Another quarter were taught by between 21% and 40% of the teachers, and 15% of the skills were taught by between 41% and 60% of the teachers. Figures 3 and 4 illustrate the percentage of teachers who reported teaching selected skills in the areas of math and language in the spring of 1983.

Construct Validity. Two types of construct validity were established for the Head Start measures: validation of the hierarchical developmental structure of each scale and validation of the assumption that each scale in the HSMB represents a unidimensional construct; that is, that each scale measures a separate construct.

Validation of the developmental structures involved application of latent class models (Bergan, 1983) to every

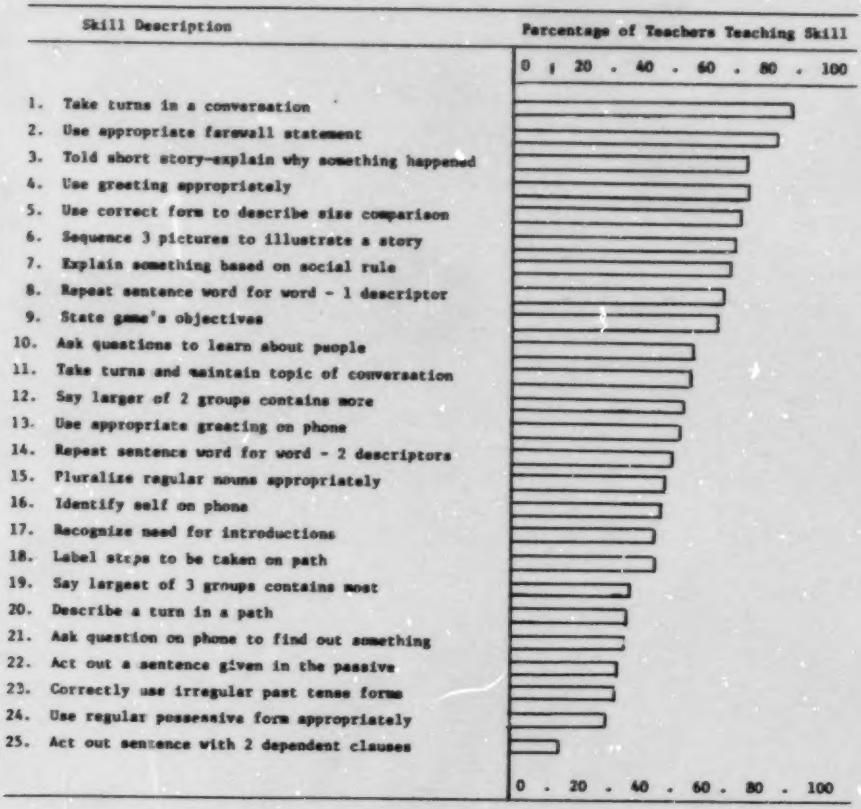

Skill Description
Percentage of Teachers Teaching Skill
0 . 20 . 40 . 60 . 80 . 100
1. Take turns in a conversation
2. Use appropriate farewell statement
3. Told short story-explain why something happened
4. Use greeting appropriately
5. Use correct form to describe size comparison
6. Sequence 3 pictures to illustrate a story
7. Explain something based on social rule
8. Repeat sentence word for word - 1 descriptor
9. State game's objectives
10. Ask questions to learn about people
11. Take turns and maintain topic of conversation
12. Say larger of 2 groups contains more
13. Use appropriate greeting on phone
14. Repeat sentence word for word - 2 descriptors
15. Pluralize regular nouns appropriately
16. Identify self on phone
17. Recognize need for introductions
18. Label steps to be taken on path
19. Say largest of 3 groups contains most
20. Describe a turn in a path
21. Ask question on phone to find out something
22. Act out a sentence given in the passive
23. Correctly use irregular past tense forms
24. Use regular possessive form appropriately
25. Act out sentence with 2 dependent clauses
0 . 20 . 40 . 60 . 80 . 100

Figure 3. Percentage of teachers who reported teaching selected language skills-Spring 1983.

BEST COPY AVAILABLE

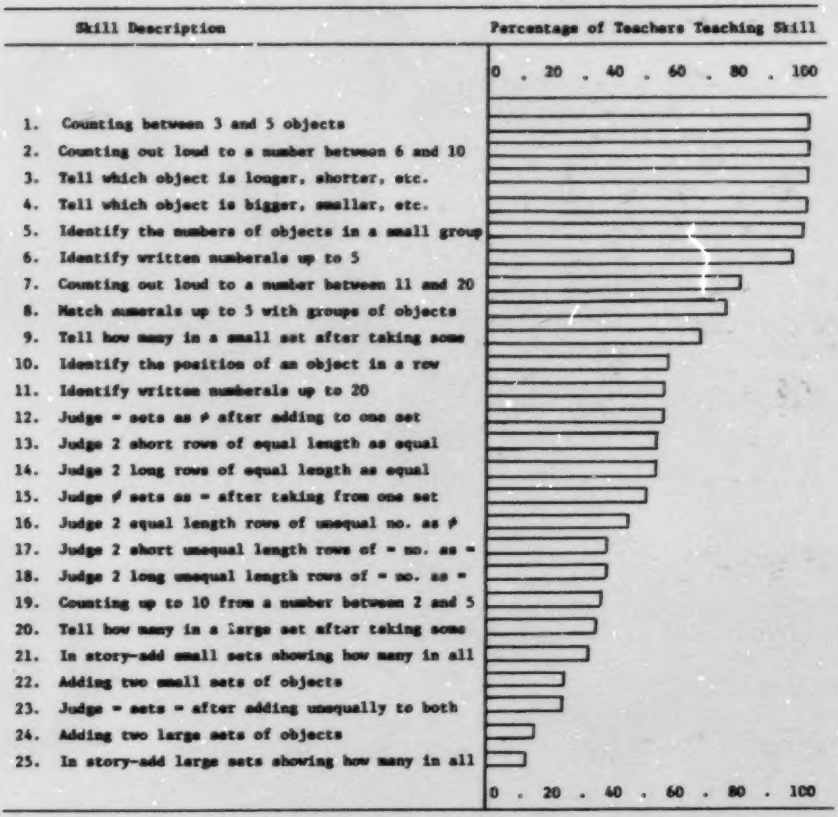

Skill Description
Percentage of Teachers Teaching Skill
0 . 20 . 40 . 60 . 80 . 100
1. Counting between 3 and 5 objects
2. Counting out loud to a number between 6 and 10
3. Tell which object is longer, shorter, etc.
4. Tell which object is bigger, smaller, etc.
5. Identify the numbers of objects in a small group
6. Identify written numberals up to 5
7. Counting out loud to a number between 11 and 20
8. Match numerals up to 5 with groups of objects
9. Tell how many in a small set after taking some
10. Identify the position of an object in a row
11. Identify written numberals up to 20
12. Judge = sets as ≠ after adding to one set
13. Judge 2 short rows of equal length as equal
14. Judge 2 long rows of equal length as equal
15. Judge ≠ sets as = after taking from one set
16. Judge 2 equal length rows of unequal no. as ≠
17. Judge 2 short unequal length rows of = no. as =
18. Judge 2 long unequal length rows of = no. as =
19. Counting up to 10 from a number between 2 and 5
20. Tell how many in a large set after taking some
21. In story-add small sets showing how many in all
22. Adding two small sets of objects
23. Judge = sets = after adding unequally to both
24. Adding two large sets of objects
25. In story-add large sets showing how many in all
0 . 20 . 40 . 60 . 80 . 100

Figure 4. Percentage of teachers who reported teaching selected math skills-Spring 1983.

BEST COPY AVAILABLE

measure to determine the ordering among tasks. Two kinds of order were examined. One involved ordering tasks by difficulty from easy to hard. The second involved prerequisite ordering. In prerequisite ordering, easier tasks are necessary to the mastery of harder tasks. Both types of ordering occur in developmental sequences. The results of the latent class analyses for the Head Start measures indicated that in the majority of instances the hypothesized sequencing of skills was confirmed by the data.

Construct validity questions related to the uniqueness of each scale involved in the assumption that each of the content areas targeted for assessment -- i.e., Language, Math, Nature and Science, Perception, Reading, and Social Development -- reflects a separate path of development. This assumption was investigated using confirmatory factor analysis (Joreskog & Sorbom, 1979). As in latent-class analysis, confirmatory factor analysis involves comparison among models. The results of the confirmatory factor analysis indicated that a model which assumed that each of the six measures in the battery would reflect a separate factor was preferred over other models examined. This model was congruent with the hypotheses underlying measures construction.

Criterion-related Validity. It is useful to assess the extent to which the HSMB relates to other measures of achievement used with young children. Other existing measures may be thought of as criterion variables to which performance on the HSMB should be related. Thus, evidence of relationships between the Head Start measures and other existing assessment devices helps to establish the criterion-related validity of the instruments.

In order to establish criterion-related validity, a sample of the children, who received the spring 1983 versions of the HSMB, were administered the Metropolitan Readiness Test (MRT) and the Preschool Inventory (PSI). The sample sizes ranged from 56 to 109 children. The Language, Math, Nature and Science, Perception, Reading, and Social Development scales correlated .10, .39*, .38*, .17, .27*, and .10 with MRT scaled scores respectively; and .50*, .56*, .66*, .41*, .71*, and .62* with PSI scores respectively (asterisks denote significant correlations). The low correlations of some of the HSMB scales with the MRT may be related to the lack of relationship between the MRT and the Head

Start curriculum. The correlations of the scales with the PSI were quite high considering that the PSI is a global measure of achievement and that the HSMB scales measure more specific content areas. These correlations are important since there is evidence that the PSI does predict later school achievement for preschool children. The findings should nevertheless be considered tentative since the sample sizes were small.

Sociolinguistic Validity. Sociolinguistic validation studies were conducted to assess the adequacy of the Spanish version of the measures for children speaking different varieties of Spanish. After fall and spring field tests, a group of reviewers were selected who were experienced with the measures, highly proficient in Spanish and English, and were speakers of the major varieties of Spanish found in the Head Start population. Each item in the Spanish version of the measures was expected to meet the following criteria: (1) the Spanish should sound natural; (2) the language should not be above the level of the children's language, i.e., not too formal or adult-like; (3) the Spanish and English versions of those items intended to be parallel should be equivalent in meaning; and (4) the variety of Spanish used should be understandable to children speaking different varieties of Spanish. The reviewers rated each item as adequate or inadequate on these criteria.

Table 4 presents a summary of their ratings. Results of the analyses of their ratings indicated that the vast majority of the items were judged to be adequate. When an item was found to be indadequate in some respect, reviewers supplied suggestions for revisions. Revisions were made based on their input as well as the input of an expert in Spanish sociolinguistics.

Table 4

Mean Number and Percent of Spanish Items Judged Adequate by Reviewers

	Natural Sounding	Child-Level Language	Equivalent Meaning	Variety Understandable
N	547	523	527	549
%	98.7	94.4	99.8	99.0
Total Items Rated	554	554	528[a]	554

[a]Items not parallel in Spanish and English were not rated.

Program Variables Associated with Achievement on the HSMB

Part of the validation of the HSMB included an examination of the influence of Head Start educational program variables on achievement assessed by means of the HSMB. Three major types of instructional program variables were examined: (1) classroom variables, (2) policy variables, and (3) background variables. Classroom variables are characteristics of a classroom that are directly controlled by the teacher in the classroom and thought to be highly related to achievement. Policy variables are program and classroom characteristics and resources that are assumed to have an indirect effect on learning by influencing the activities that take place in the classroom. Policy variables are amenable to alteration by administrators but are not generally subject to direct manipulation by classroom teachers and aides. Background variables are characteristics of children and their families.

Learning Opportunities. The first classroom variable examined was the amount of time devoted to providing learning opportunities. Teachers participating in the project were asked to indicate the amount of time in their daily schedules that they spent providing teacher initiated

learning opportunities within specific content areas. The amount of time allocated to such learning opportunities was found to be related to student achievement in the areas of math, nature and science, social development, and perception. This means that teachers who spent more time providing learning opportunities had students who demonstrated higher levels of achievement on the HSMB by the end of the program.

Teacher Knowledge of Children's Skills. The amount of knowledge that teachers have about what their students know and don't know is related to achievement. Each teacher involved in the project used a **Planning Guide** to indicate for each child those skills that had been mastered, those skills that had not been mastered, and those skills that had not been taught. This information was compared with the actual performance of each child on the HSMB in order to obtain an index of "teacher knowledge of children's skills." The results indicated that the knowledge variable was related to achievement on all six scales in the Battery. This means that teachers who were sensitive to a child's level of skill in a given content area had a greater impact on child achievement than teachers who lacked sensitivity to skill level. This finding supports the view that a teacher who knows what skills a child possesses and what skills a child does not possess is in a better position to help the child than a teacher who does not have that kind of information.

Individualization. Data obtained from the **Planning Guide** were used to determine level of difficulty of the skills teachers reported they were teaching to each child. Latent trait estimates of item difficulty were used to code the difficulty levels of each task reflected in the **Planning Guide.** An instructional difficulty index was computed by averaging the difficulty levels for the skills taught to each child. The difficulty level of skills that were taught was calculated for individual children as well as for classes as a whole. In many cases the standard deviation for the class was zero. This indicates that all children in such classes were being provided with the same learning opportunities. In Reading, Level I skills, for example, 80% of the teachers reported teaching the same skills to all children, while 20% provided for some individualization. For Reading, Level II skills,

85% reported teaching the same skills to all children and the other 15% reported some individualization.

The policy variables examined in the project fall into three broad categories: program exposure, classroom composition, and teacher training and qualifications. Each of these is described below.

Program Exposure. Program exposure was defined as the length of the instructional day, the length of the instructional year, and the number of days during the instructional year that each child had actually been in attendance. Data for these three exposure variables were obtained from attendance forms and teachers' schedules.

The policy variable that was found to be consistently associated with achievement on the HSMB was program exposure expressed as total days present at Head Start. The number of days that each child had been in attendance was related to achievement on the Language, Math, Nature and Science, and Perception Scales. This means that those children who were rarely absent showed higher levels of achievement on these four measures by the end of the school year than those children who were absent often. Reading was again one of the Scales for which no effect was found; however, the lack of emphasis on teaching reading skills provides a plausible explanation for this finding. The other Scale for which an effect was not present was the Social Development Scale. There is no apparent reason for the observed lack of effect for this scale.

No significant relationship was found between the length of the day and performance on the HSMB nor between the length of the year and performance on the HSMB. A plausible explanation for the lack of observed relationship between the length of the day and achievement may have to do with when learning opportunities are provided to students. The number of learning opportunities was found to be an important prediction of higher levels of achievement on the HSMB, but it may be that there is not a great deal of difference between the amount of teaching that occurs in half-day and whole-day programs. A possible explanation for the lack of significant relationship between performance on the HSMB and length of the year is that there was not a great deal of variability in the length of the year across programs. There may have been more variation in

the number of days a child was present (not absent) than in the total days of the program.

Classroom Composition. Information on class size and staff/child ratio were gathered in order to examine the relationship of these variables to performance on the HSMB. Classroom observations were conducted at four different times during the year. Means were obtained across the four observations. The mean class size was found to be about 17 with one teacher and one aide present on the average. No significant relationship was found between class size and performance on the measures, nor was a relationship found between staff/child ratio and achievement. Although class size has been found to be related to achievement in other studies (see, for example, Smith & Spence, 1980), the most plausible explanation for the lack of relationship in this study seems to be the lack of variability in class size and staff/child ratio among the classrooms in the sample. This issue is again being examined in the spring 1984 pilot study described in Chapter IV.

Teacher Characteristics. Information on teacher qualifications examined in the Head Start Measures Project included amount of education, degrees and certifications obtained, and amount of prior experience in teaching. Data on these variables were obtained from a classroom staff questionnaire. Teaching experience was defined in terms of the number of years of prior teaching activity including experience as a Head Start teacher, a Head Start aide, a Head Start volunteer, a preschool teacher, a preschool aide, public school teacher, private/parochial school teacher, home visitor, and day care classroom staff member. Information on teacher training included amount of CDA training and amount of Head Start inservice training during the current year. CDA training was calculated in months and inservice training was given in days.

Few relationships were found between teacher characteristics and spring achievement levels. Amount of Head Start inservice training (measured in days) was related to spring achievement in Language and Reading, while amount of CDA training (calculated in months) and possession of a CDA credential were related to spring achievement on the Perception Scale. The general lack of relationship between teacher training and performance on the HSMB may be due to the fact that data were not gathered on training related to specific

content areas. The relationship between training in content areas related to the HSMB and children's achievement on the HSMB is being examined in the spring 1984 pilot implementation study.

Background Variables. A number of background variables were examined along with policy and classroom instructional variables in the determination of program effects. Age, recorded in months, was included as an index of growth that could be used to separate noninstructional from instructional influences on development (Bryk, Strenio, & Weisberg, 1980). Other background variables included socioeconomic variables and family background variables shown to affect achievement (Bergan & Henderson, 1979). The socioeconomic variables included the primary family provider's occupation and income. These variables were obtained from a family background data sheet. The provider's occupation was recorded using Duncan's SEI scale. The family variables involved the mother's education and the number of siblings and parents in the family. This information was obtained from a family questionnaire.

Of those background variables examined in the Head Start Measures Project, the only variable that was found to be related to achievement was age. Based on the child-level status score analysis, age was significantly related tc achievement on all six scales. Age influences spring achievement in two ways. It affects spring achievement indirectly by influencing fall achievement which in turn affects spring performance. In addition, it has a direct influence on spring achievement. This finding shows that measures performance is influenced by developmental factors associated with age as well as by participation in Head Start.

IV

DISSEMINATION AND UTILIZATION PILOT STUDY

In February of 1984 a six-month project was initiated to pilot the dissemination and use of the Head Start Measures Battery prior to a broad scale implementation planned for the fall of 1984. The purpose of the project is to develop and evaluate training materials, produce the HSMB test manuals and manipulable materials, train Head Start personnel in measures use and interpretation, assess measures administration feasibility, and analyze data on child performance and program variables. This chapter describes the design of the study, the development of materials and computer programs, the sample of sites involved, the training activities, data collection procedures, and the planned evaluation of the dissemination effort.

Study Design

Two implementation strategies are being examined: (1) the Head Start Staff Model and (2) the Site Manager Model. Thirty programs were selected to participate and were assigned to one of the two models. Under the Head Start staff model, the HSMB was implemented by Head Start staff. An Educational Coordinator from each program was first trained in the HSMB system. He or she then selected testers from among existing Head Start staff, trained them, and supervised the testing. Under the Site Manager Model, an outside person was hired by the Head Start program to manage the testing activities. The Site Manager selected and hired testers, trained them, and supervised data collection. Late in the spring the Educational Coordinators from the Site Manager Model sites were invited to training sessions so that they could become familiar with the path-referenced assessment system.

All testing was completed by early June and HSMB data, as well as data on selected program variables, were sent to the University of Arizona for processing. A series

of six meetings were scheduled for late July to discuss the interpretation of results of the HSMB and how results might be used to improve the educational component in Head Start programs and classrooms. A final report will be available in August and will present an evaluation of the training materials, the training process, the data collection procedures, and a description of dissemination obstacles and successes encountered under each of the two models.

Development and Production

The six scales of the HSMB have been developed and continually refined since 1981. The pilot phase involved an additional set of revisions and the professional printing of 125 copies of the measures.

Four manuals were produced in the winter of 1984. The **Examiner's Manual** describes the content of the HSMB, specifies administration procedures, and explains how to interpret scores. The **Data Collection Training Manual** was developed for use by Educational Coordinators (and Site Managers) in training data collectors to administer the measures. It contains a suggested training plan and all the information necessary to adequately monitor the testing process. The **Technical Manual** describes the psychometric properties of the measures including validity, reliability, and item information. It also describes procedures employed to avoid and eliminate linguistic and cultural bias. A manual for use in interpreting test results and using the information to evaluate and improve the educational component is also being developed based on feedback from Head Start personnel participating in the pilot study.

In addition to the manuals, a series of 1/2" video cassette tapes were developed and produced for use in training sessions. They include a training module demonstrating administration procedures for each of the six scales of the HSMB.

The development phase of the study also involves the writing of computer programs for use in scoring the tests and producing a series of reports for both individual students and classes. Computer programs have also been written for item banking. The item banking technology

allows items to be placed on a common scale. This is necessary for a measures system to be adaptable to local program needs.

Participating Sites and Children

Twenty-eight Head Start programs participated in the dissemination and utilization pilot study. Participating programs were selected from among programs that volunteered and that were willing to commit the time and resources necessary to carry out the pilot. The 28 programs represented all of the ten Head Start regions. They also represented a range of program sizes in both urban and rural settings. Twelve of the sites have Spanish-speaking children. It was considered important to include a sufficient number of Spanish-speaking children so that procedures for determining whether to administer tests in English, Spanish, or bilingually could be assessed. Programs were assigned, randomly in most cases, to either the Head Start Model or the Site Manager Model. Within each site 30 children were selected at random to be tested.

Implementation

Following a week long training session, Educational Coordinators and Site Managers selected and trained data collectors. They set up testing schedules and supervised the process of data collection to ensure that procedures for proper administration and scoring were followed. In addition to administering the HSMB, each participating program collected information on selected program variables including: class schedules, attendance, classroom composition, teacher characteristics, and family income and occupation. Planning Guides were filled out by participating teachers. For each child assessed with the HSMB, the teachers indicated in the Planning Guide what skills had been taught to the child, what skills the teacher felt had been mastered by the child, and what skills had not yet been mastered. Data collection activities were completed by early June and data were sent to the University of Arizona for processing.

Evaluation

The evaluation is currently in progress and the results will be provided in a separate report. The central purpose of the evaluation activities is to determine the effectiveness of the training, the effectiveness of each of the manuals comprising the measures package, and of the implementation of the HSMB system under conditions in which measures administration was directed by local Head Start personnel and under conditions in which measures administration was directed by outside personnel. Another major purpose is to assess the administrative feasibility of the measures and to make changes as appropriate. A subordinate purpose is to provide additional information on program effects and measures characteristics identified in the measures development phase of the project. Evaluation activities for the current phase fall into four categories: evaluation of training, evaluation of system implementation, evaluation of program effects, and evaluation of measures characteristics.

Evaluation of training will focus both on the quality of the training experience and on the effects of training on project participants. Evaluation data were collected in the training workshops through the use of questionnaires. The evaluation will provide feedback for refining training materials and procedures prior to their dissemination on a broader scale.

The evaluation of training effects was carried out to determine the extent to which participants gained the necessary skills to implement the HSMB in local Head Starts. Data on assessment skills were gathered using shadow scoring procedures and a data collection monitoring form both during training and on site.

The evaluation of system implementation is being carried out to determine the extent to which the skills acquired during training are implemented by participants in local Head Starts. The evaluation of system implementation will focus on training in local sites, data collection, the interpretation of results of testing, and the use the results of the measures for improving educational plans. A variety of instruments and procedures are being used to assess system implementation. Site Managers and

Educational Coordinators used the shadow scoring and monitoring methods forms to determine the adequacy of data collection. Quality control procedures designed during the measures development phase of the project were implemented at the University of Arizona to serve as a further check on the adequacy of data collection. The administration of the **Planning Guide** was evaluated by T/TA providers. Interpretation skills and planning skills will be evaluated from data obtained during group meetings held in July.

Evaluation planned for the project includes the examination of program effects using test performance data, **Planning Guides**, and program variable instruments. The program variable instruments will provide information that can be used to reexamine the relationship between program characteristics and performance on the HSMB.

The administration of the measures during the spring of 1984 affords the opportunity to update information on their psychometric characteristics. Moreover, the examination of the properties of the measures affords a means for testing the implementation of the item banking technology being developed for the project. Examination of item statistics including item difficulty and discrimination will be particularly important. The properties of test items may change over time. In order to maintain items that can be used to form scales accurately reflecting children's abilities, it is useful to update item statistics whenever the measures are administered on a broad scale. Latent trait techniques employed during measures development will be used to update information on item characteristics during this phase of the project.

Although results are not yet available from the pilot study, informal feedback has already been used to revise the battery. Each of the six scales has been shortened. Items that were least relevant to the Head Start Program were deleted. In addition, the **Examiner's Manual** was simplified. These changes will facilitate the use of the HSMB Head Start programs in the fall of 1984.

References

Bergan, J.R. (1983). Latent-class models in educational research. In E.W. Gordon (Ed.), **Review of Research in Education**, Washington, D.C.: American Educational Research Association, 10, 305-360.

Bergan, J.R. (1981). Path referenced assessment in school psychology. In T.R. Kratochwill (Ed.), **Advances in School Psychology**. (Vol. 1). New Jersey: Lawrence Erlbaum Associates.

Bergan, J.R., Stone, C.A., & Feld, J.K. (1984). Rule replacement in the development of basic number skills. **Journal of Educational Psychology**, 76, 289-299.

Bergan, J.R., Anderson, D.O., Feld, J.K., Henderson, R.W., Johnson, D.M., Lane, S., Mott, S., Parra, E., Robinson, L., Stone, C.A., & Swarner, J. (1984). **Head Start Measures Final Report.** Prepared for the Department of Health & Human Services under Contract No. HHS-105-81-C-008.

Grottberg (1969). **Review of Research: 1965-1969.** Washington, D.C.: Project Head Start, U.S. Office of Economic Opportunity.

Hambleton, R.K. & Murray, L.N. (1982). **Some goodness of fit investigations for item response models.** Amherst, MA: University of Massachusetts, Laboratory of Psychometric and Evaluative Research Report, No. 129.

Head Start Program Performance Standards. (1975). Office of Child Development, Department of Health, Education, and Welfare.

Joreskog, K.G., & Sorbom, D. (1979). **Advances in factor analysis and structural equation models.** Cambridge, Mass: ABT.

Lord, F.M. (1980). **Applications of item response theory to practical testing problems.** Hillsdale, N.J.: Lawrence Erlbaum Associates, Inc.

Smith, A. & Spence, C. (1980). National day care study:- Optimizing the day care environment. American Journal of Orthopsychiatry, 50 (4), 718-722.

Project Director:	
John R. Bergan	University of Arizona
Government Project Officer:	
Allen N. Smith	Administration for Children, Youth and Families

Project Managers

Jason K. Feld	
Donna M. Johnson	
Kathleen Silvers	
Clement A. Stone	University of Arizona

Authors of the HSMB:

Language Scale:

Joyce C. Swarner	
John R. Bergan	
Stacey E. Mott	
Adam Schmaps	
Teresa K. Anderson	
Elena Parra	University of Arizona

Math Scale:

John R. Bergan	
Jason K. Feld	
Clement A. Stone	
Adam Schmaps	
Joyce C. Swarner	
Teresa K. Anderson	
Elena Parra	University of Arizona

Nature and Science Scale:

Ronald W. Henderson	University of California, Santa Cruz

Perception Scale:

Rosemary A. Rosser	University of Arizona

Reading Scale:

Joy L. Hestand	University of California, Santa Cruz

Social Development Scale:

Sadie Grimmett	Indiana University

For more information contact:

Dr. Allen N. Smith
Administration for Children, Youth and Families
400 Sixth Street, S.W., Room 5143
Donohoe Bldg.
Washington, D.C. 20013
(202) 755-7724

Center for Educational Evaluation & Measurement
College of Education
The University of Arizona
Tucson, Arizona 85721
(602) 621-7859